International Journal of Hybrid Information Technology
Vol. 9, No.9 (2016), pp. 423-432
http://dx.doi.org/10.14257/ijhit.2016.9.9.38

Wavelines: Visualization Method for Comparative Analysis of Time Series Data

Ruijun Liu[1,4], Ningning Liu[2], Yi Chen[1,4], Yunfang Zhao[1] and Yang Xu[3,*]

[1]*Beijing Technology and Business University, Beijing 100048, China*
[2]*University of International Business and Economics, Beijing 100029, China*
[3]*Peking University, Beijing 100871, China*
[4]*Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing 100048, China*
Yang Xu, yang.xu@pku.edu.cn

Abstract

Time series data has characteristics of huge differences among date values, non-continuity and long time span. Cloudlines layout algorithm can overcome the shortcoming of growing data over time, for the data is non-continuous, cyclical and greatly different in value, and the display is not ideal for space waste and form of a single attribute value data in Cloudline. In the thesis, we put forward the Cloudlines Map Visualization Analysis Methods according to the characteristics and shortcomings of time series data. The methods include data values using a logarithmic transformation, semi-ellipse superimposed waveform, transparency distinguishing the density, waveform smooth transition and unequal axes. The WaveLines diagram applied to pesticide residue data can help users effectively to analysis and compare data trends of multiple dimensions over time.

Keywords: *time-series data; Wavelines; visual analysis; pesticide residues*

1. Introduction

There are always new data constantly generated, and each new data has the time dimension attribute value. It is similar to the visual analysis of the real estate data, import and export data [8] and social network data [1] and so on, which they all have obvious attribute of time dimension.

Time-series data is likely to be complex, often they not only have simply single time dimensional values, at the same time they may also have multiple values relating to time. Between time-series data of the same type, there always exists associations which need to do combine comparison or analysis.

For the time-series data, analyzing from the periodic, the sampling time may be periodic or aperiodic. For the periodic time-series data, we can infer some rules from the periodic change to analysis. For the aperiodic time-series data, there often exists many limits in the application of visualization method. One of the most difficult problem is that the time-series span is so long which needs large amounts of data analysis. Therefore, doing visual analysis for the aperiodic time-series data, is a difficult problem now.

Received date: 2016-04
Supported by the open funding project of State Key Laboratory of Virtual Reality Technology and Systems, Beihang University (Grant No. BUAA-VR-16KF-18)
Correspondence should be addressed to XU Yang, E-mail: yang.xu@pku.edu.cn

ISSN: 1738-9968 IJHIT

The visualization of the time-series data has been in continuous research in recent years. Because of the time-series data are often so huge, the description accuracy of a single moment will be have some lost. In this case, people are more likely to focus on the following aspects in analyzing time-series data: 1) Changes of the historical data; 2) The contrast between the new data and the history data; 3) The change trends of the future data.

There exists correlation within time-series data. For example, the news, for a certain event, from different sources have correlative report about the certain event. In view of this situation, we may need to do some contrast and analysis for the time-series data coming from multiple sources, and analyzing the relationship and the characteristics between the data. Krstajic [2] and others proposed Cloudlines in 2011, applying to news data visualization analysis. However, Cloudlines exists 3 following drawbacks: 1) the display effect are axisymmetric. The information displayed on the negative half axis of the vertical axis are exactly the same with the positive semi axis which wasting the display space. 2) When the data is not continuous and there are sparse regions, the display effect will exist in the corresponding blank area. 3) When it comes to analysis a single moment, the value of data expressed only with color. In the light of the above disadvantages, this paper presents a visualization method called WaveLines Chart, which is suitable to analyze the aperiodic data and make comparison analysis.

The detection results of pesticide residue data, which we call it pesticide residues as abbreviation, are known to have detection and location, test time and test results, the kinds of pesticides, agricultural products, pesticide detection result value, pesticides exceeded the limit value and so on, totally 13 attribute values. From the categories of the properties, we can see that pesticide residue data has obvious attribute values of time dimension which is suitable for the visualization method in light of time-series data to analyze and research. Observing the specific pesticide residue data, it is difficult to find out the periodicity. In the process of using pesticide, there exists the following 5 situations: 1) Pesticides may use as collocation; 2) Different pesticides may have the same effect; 3) Different agricultural products may use the same kind of pesticides; 4) The same kind of agricultural products may use different kinds of pesticide; 5) The differences between pesticide residues may be very large. Therefore, in view of the above situation, we need to do comparative analysis of the pesticide residues data.

2. Related Work

The visualization method of time-series has always been an important research direction in visual field. Aigner and others [3] make some simple analysis to the visualization method of time-series. The simplest way to display is the Line Graph [15]. Playfair and others [8] use a line chart to analyze the relationship between imports, export, and price and so on with time. Wattenberg and others [9] use the stacking technology based on the Line Graph, which is used at the display contrast of multiple time-series data at the same time. ThemeRiver [11], applying to the text visualization analysis, also applies stacking technology.

Saito and others [13] propose the Horizon Graph using the depth of color to express the variance ratio of the time-series data. While Heer and others [14] make the further development of this technology. In order to analysis the time-series data better, Perin and others [16] combine the Line Graph and the Horizon Graph through interaction technology.

Besides the application of the horizontal time axis, Kehrer and others [12] use the cyclic time axis, using the spiral ring structure to represent the beginning of time, and using the information of width and color and so on to represent the attribute value of the time-series data. While Claessen and others [10] pull the time element in the parallel coordinates. Fuchs and others [7] make contrast analysis between the visualization

method of linear, tape, disk and star graph in the application of the analysis of time-series data.

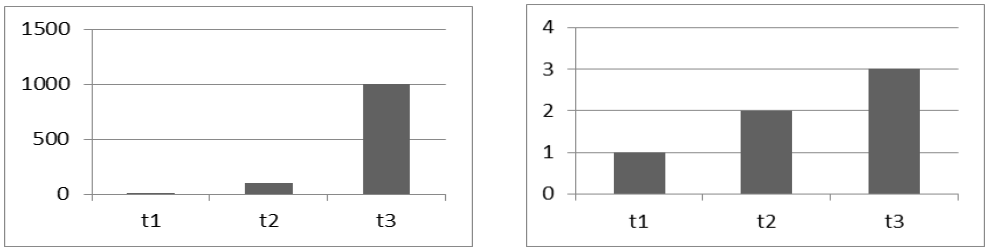
In view of the contrast between time-series data, Javed and others [6] put forward the Braided Graph. Hao and others [4] conduct large-scale analysis of time-series data in view of sampling data with different time, using different display classes. McLachlan and others [5] combine multiple visualization technology and make use of the interaction technology to do some linkage between technologies and design a visualization analysis system for time-series data.

3. Wavelines Chart

For the aperiodic time-series data, at some point, may not exist the corresponding value. Therefore, when using time axis related visualization methods to display, it may easily cause blank phenomenon in a certain period of time, which will influence the rate of screen using. When it comes to a large amount of data, if they are displayed in the horizontal axis with the pixel 1000, the size of pixels of each time data can occupy is 4. After average distribution, the 1000 pixels can only display 250 moments of the data. However, the span of time-series data often very long, and the amount of data increase as time going by. In the limited screen and in the situation not using any interactive technologies, it can only display the limited data which is difficult to fulfill the visualization demands of time-series data. Within a view, it's difficult to display the whole time-series data, thus comes the macro analysis. If you use transparency to distinguish the overlapping area between different moments, the displaying number could improve, however, the data at a single moment would be difficult to distinguish. In that case we can only observe the data in overall situation.

2.1. Logarithm Value

Within the time-series data, the data values at different moments are often not the same. However, when analyzing data from the whole situation of time-series data, the difference value between the historical data with the new data may be very large. For example, in the pesticide residues data of strawberry, the detectable amount of Omethoate pesticide is 0.0512mg in September 13th, the detectable amount change to 0.0014mg in September 14th and 0.0183mg in September 15th, during this period there are 51.1 times of the difference. Therefore, if we use the data value directly to do visualization layout, it would lead to the smaller value of those data which can't be clearly displayed and recognized on the time axis. In view of this situation, in this paper, we use the logarithmic form to converse data value.



(a) Before taking the logarithm (b) After taking the logarithm

Figure 1. The Difference in Displaying Before and After Taking Logarithm

L is a User defined coefficient, if the value of the time-series data that corresponding to the I time is v_i, taking $w_i = log_L^{vi}$ to do the logarithmic transformation where w_i is the value of visualization method. To the time-series V, for example, which the value v_1 in

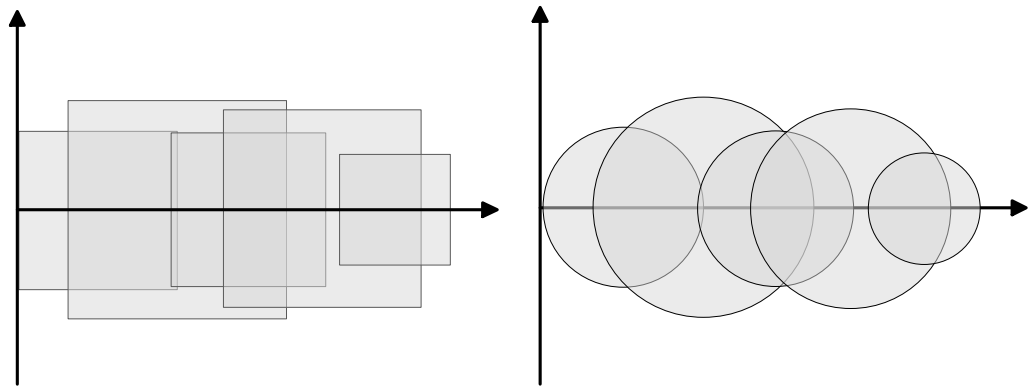
time t_1 is 10, the value v_2 in t_2 is 100 and the value v_3 in t_3 is 100. If we use the histogram in Figure 1 to display, the visual effect in time *t1* would be very bad.

If the displaying area that corresponding to time t_2 occupy a height of 100 pixel, the displaying area that corresponding to time t_1 only occupy a height of 1 pixel. We suppose *L=10*, after a logarithmic transformation, in which the value tends to be 1 at time $t_{1,2}$ at time t_2 and 3 at time t_3, the largest difference of 3 times. At this time, it can be displayed clearly on the screen no matter at time t_1 or t_2.

2.2. Semi Ellipse

Because the spacers between each moment on the time axis are fixed, it is advantageous to distinguish different moments following the sequence of the time axis. When using a square or a circle as a data displaying form of each moment, if we use radius or area to represent value, it would cause the irregular blank or the irregular overlap phenomenon, as is shown in Figure 2(a) and Figure 2(b). The irregular overlap and the blank area may influence user to judge data. In view of this problem, the WaveLines Chart would use ellipse to represent data of each single time.

Supposing the ellipse is consisted of the long axis r_l and the short axis r_s, the coordinate axis is Descartes coordinate system, then the method of calculation is: defaulting the intervals between the x coordinate are equal to the short axis r_s, the height of the y axis is n, the data value is equal to the logarithm at that moment, as is shown in Figure 3. When $2*r_s$ is greater than the interval between the x axis coordinate, there would exist overlap phenomenon. Using rectangle to resolve the irregular overlap and the blank area has the same effect. But the elliptical method is more smoothly, and the display method is more consistent with the form described in section 2.4, which is in line with the cognitive of users.

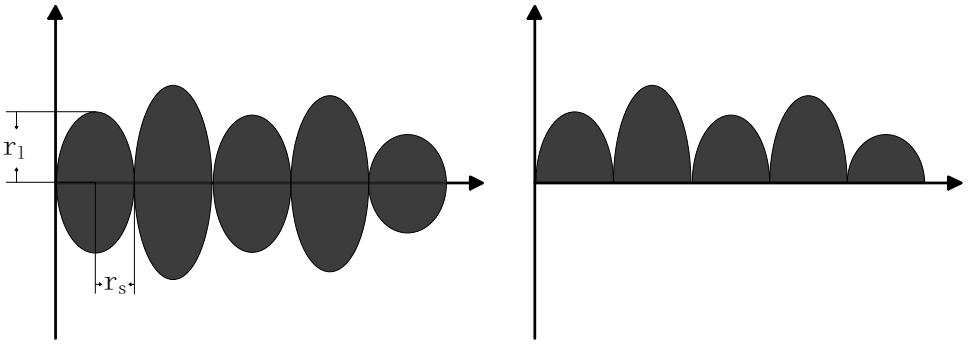


(a) Square (b) Circle

Figure 2. The 2 Kinds of Shape Display Effect Chart

When using ellipse, hiding the negative half axis of y axis does not affect the data analysis results and at the same time it saves half pixel space and conducive to display more content in limited space. Using a semi elliptical manner to display will have the same display effect and the occupation of the pixel area is smaller. It can use the negative axis of y axis to display related data doing contrast analysis.

When displaying a long time data and there the number is huge, it would optimize the displaying effect using the transparency mentioned in section 2.3 and the smooth transition mentioned in section 2.4. The displaying effect after optimizing is shown in Figure 4b.



(a)Ellipse (b)Semi Ellipse

Figure 3. Description of Ellipse Display

2.3. Transparency

Displaying the whole content of the data on a screen then analyzing the change trend overall needs to keep the display effect and the amount of data as far as possible. In order to avoid the situation that the average pixel weight of single moment can't be recognized by human's eyes, we increase the amount of the data that can be displayed.

When displaying a large amount of data, there comes with the phenomenon of overlap. When the data is intensive and if we do nothing to optimize them, it is difficult to find the density of the overlap data. In order to reduce the influence the overlap may bring to us, we use transparency to distinguish the density of the overlap area in the visualization effect.

If the display pixel upper region is p, the amount of the time-series that needs to be displayed is q, each time the data possessing the pixel width is w, when $q*w>p$, the displaying area will exist the overlapping phenomenon, which the maximum value of the overlapping area is $cm=ceil（q*w/p）$. Supposing the transparency is 0% where the overlapping region number is $(1-ci/cm)*100\%$, the transparency corresponding to the no data display region is 100%.

The transparency is $(1-ci/cm)*100\%$ in which the number of the overlapping region is ci. Through transparency, we can increase the amount of the time-series data, we can also find the frequent degree in different time zones.

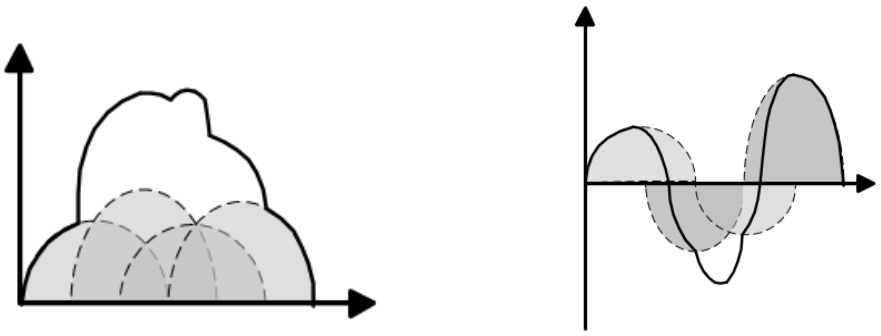
2.4. Smooth Transition

No matter using which shape pattern to show the time-series data, because the value is not the same size, data at different times combined into overlapping region will in irregular pattern. If distinguishing only through the transparency change, then for the time-series data of long time span there will exist large number of irregular pattern consisted of transparency. A limited display region not only can't express clearly the value of each mount, it can't express even the change situation of each mount clearly.

Under the situation that time-series data grows as time growing, comparing with the single data situation, the more important thing is to find the change trend through visualization methods. When using the semi ellipse, it can't express the time-series data accurately in a limited space visually, it will use the wave accumulation method to optimize the results, making the whole display effect more smoothly.

Through accumulating the value of overlapping regions, it integrates the semicircle to a curve, as is shown in Figure 4(a). For a curve after a smooth transition, transparency will be classified according to the level regional of the overlap region. If the intersection of the semicircle and the x axis of all moments is n, then set the intersection as $\{x_1, x_2, x_3, \ldots\ldots, x_i, x_{i+1}, \ldots\ldots, x_n\}$, within the arbitrary regional scope $[x_i, x_{i+1}]$, using the transparency which is corresponding to the highest overlapping number as the transparency of the region.

When using negative axle to show periodic data for comparison, subtraction can be used to make a smooth transition, such as Figure 4(b).



(a) Superposition (b) Subtraction

Figure 4. Schematic Diagram of a Smooth Transition

2.5. Unequal Axial Width

Time-series data in the time axis may not be continuous, there may exist a data blank situation at some moments or in some period of time. And besides, in the comparison of multiple data, there may exist a phenomenon that the most data source don't have values but very few data. In view of these outliers, users may not want to observe. In view of the two situations mentioned above, we use the threshold method to divide the unequal axial width. The division way reference two factors, in view of some moment *t*, considering the continuity *Ft* and the coincidence rate *Gt*, the calculation method of the threshold E_t is $E_t = F_t + G_t$.

If there exists data at n different moments, the continuity calculation formula is $F_t = \sum_1^n f_i$ among which f_i is the continuity corresponding to the data source *i*, as is shown in formula (1), and *k* is a coefficient that the user defined.

$$F_i = \begin{cases} 1 \text{ The values corresponding to the nearby } k \text{ moments are incompletely 0.} \\ 0 \text{ The values corresponding to the nearby } k \text{ moments are completely 0.} \end{cases} \quad (1)$$

$G_t = m$, m is the data source number that having values at the moment *t*.

We set the interval between moments to the class *K* of wide spacing and the class *Z* of narrow spacing. In view of the user-defined threshold E_f, when Et is bigger than the threshold E_f, judging the moment as class *K*; when E_t is smaller than the threshold E_f, judging the moment as class *B*. Through the user-defined proportion of the moments corresponding to class *K* and class *Z*, we will realize the proportional distribution of the time axis. And the moments of class *K* and class *Z* will distribute the time axis length on average that possessed by the class.

3. Experiments

In order to test the visual effect of the Mountain Line Graph, this paper simulates the data which is based on the real pesticide residues, using the Mountain Line Graph to do visualization analysis. The display of pesticides data of different names includes the daily total amount of pesticide residues detection and the corresponding number exceeds the standard in vegetables & fruits, as is shown in Figure 5. The total amount of pesticide residues detention is displayed by a smooth transition mode, using logarithm as value, taking *L* to 5; using semi ellipse to display the number exceed the standard; displaying the value as the original value; rs is 4 times of the time axis interval; the threshold *Et=1*;

using unequal axial width, and *K:Z=5:1;* one scale of the time axis represents one day. Figure 5 is the amount of pesticide Dicofol detected in 2013 compared with the 2014, using subtraction waveform to display the chart. We regard weeks as the horizontal alignment of standards between different years.

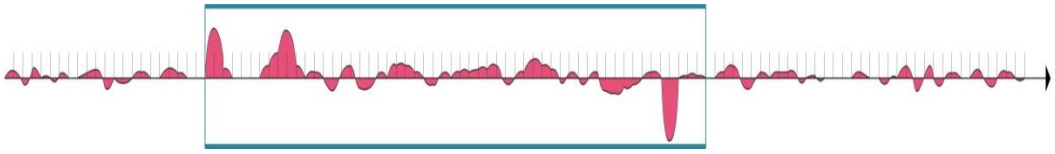


Figure 5. The Application of Pesticide Dicofol Residue Data of Strawberry in 2013 And 2014

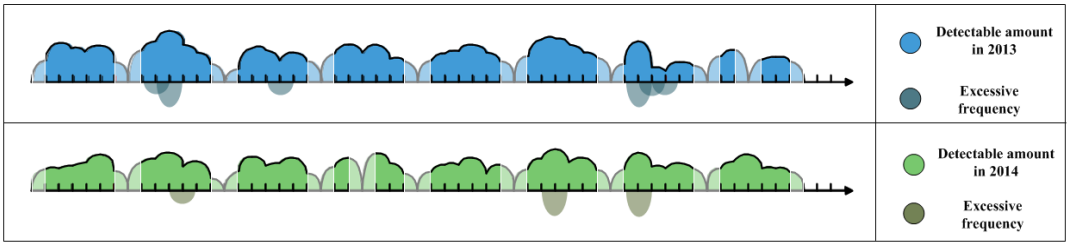


Figure 6. Attention to the Data from March Last for 8 Weeks

From Figure 5, most of the time, the data fluctuate gentler, the total amount of pesticide residues is not much difference. Generally speaking, the amount of pesticide detection in 2014 is more than pesticide residues detection in 2013, we can infer that pesticides Dicofol tends to be used in standardized way and the residual amount after using has decreased. But there are individual excessive amounts of pesticides which were detected in 2013 and 2014, appearing as a small peak in Figure 5. In this case, we select one obvious area stretching and splitting into 2013 and 2014 two patterns to analyze. Figure 6 is generated by selecting period of time of 8 weeks after the start of March.

Figure 6, using color to distinguish the year of pesticide residues detection and displayed excessive frequency. Looking from the overall, pesticide residue situation is widespread but fewer exceeding the standard, which the day passing rate is more than 98%. As the number of detected samples every day is not the same, the residues of big amount may not exist the exceeding standard situation, however the residues of small amount may exist the exceeding standard, so there is no rules to follow.

Every 5 time intervals there exists the area which the opacity is 50% and the image shows the sunken that indicating there is no detecting result data at this moment. As dividing the axis width of the no data areas unequally using the rate of 5:1, then, when displaying the no data areas that possessing one moment represent two days. For the area of 50% transparency, we can find the overall data having a cycle of 7 days. Generally there exists pesticide residue data on working day, and not exists on the nonworking day.

In Figure 6 of 2013, the presence of pesticide residues detection value of the seventh months displays an apparent peak. Compared to the superscalar which it has exceeding cases, and the pesticide detection values presence large data. The subsequent in two days detectable amount is not large, the situation did not change. These days require a detailed analysis and have to strengthen monitoring.

When comes to analyze the aperiodic transparent part, there doesn't exist the corresponding pesticides residues detection on Wednesday of the eighth week and Wednesday of the fourth week. Because of the pesticide residue detection result is accurate to 0.0001 mg, we can infer that on Wednesday of the eighth week and Wednesday of the fourth week the strawberry samples are not using the pesticide Dicofol respectively.

4. Conclusion

In this paper, through the comparison of temporal existing time-series data visualization methods with the Cloudlines, we summarize the Cloudlines' disadvantages, proposing a visualization method that is more flexible and more applicable to time-series data — the WaveLines Graph. The method use logarithm to replace, semi ellipse to display, transparency to optimize, and do intensive data smoothing computation and refer the unequal axis width and so on, make visual presentation of time-series data, which is conductive to analysis the overall trend of the data.

In this paper, we use the WaveLines Graph into the analysis of the pesticide residues data, and get some analysis results. In the future, it can be popularized and applied to other area's data to do visual analysis such as the flow of people, the blog data and the weather data and so on. Although we use a semi elliptical structure to do optimization, in the same condition, using WaveLines Graph to display data, the amount of data is 2 times of Cloudlines. But there is still an optimization space. We can learn from some other technologies such as the Braided Graph 0, using the method into the WaveLines Graph can increase the amount of data's parallel analysis and the form of visualization.

References

[1] X. Yuan, L. Che and Y. Hu, "Intelligent Graph Layout Using Many Users' Input", Visualization and Computer Graphics, IEEE Transactions on, vol. 18, no. 12, **(2012)**, pp. 2699-2708.
[2] M. Krstajic, E. Bertini and D. A. Keimm, "Cloudlines: Compact display of event episodes in multiple time-series", Visualization and Computer Graphics, IEEE Transactions on, vol. 17, no. 12, **(2011)**, pp. 2432-2439.
[3] W. Aigner, S. Miksch, W. Muller, H. Schumann and C. Tominski, "Visualizing time-oriented data–A systematic view", Computers & Graphics, vol. 31, no. 3, **(2007)**, pp. 401–409.
[4] M. Hao, D. Keim, U. Dayal and T. Schreck, "Multi-resolution techniques for visual exploration of large time-series data", In Eurographics/IEEE VGTC Symposium on Visualization, Norrkoeping, **(2007)**.
[5] P. McLachlan, T. Munzner, E. Koutsofios and S. North, "LiveRAC: interactive visual exploration of system management time-series data", In Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, ACM, **(2008)**, pp. 1483–1492.
[6] W. Javed, B. McDonnel and N. Elmqvist, "Graphical Perception of Multiple Time Series", Visualization and Computer Graphics, IEEE Transactions on, vol. 16, no. 6, **(2010)**, pp. 927–934.
[7] J. Fuchs, F. Fischer and F. Mansmann, "Evaluation of alternative glyph designs for time series data in a small multiple setting", Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, **(2013)**, pp. 3237-3246.
[8] W. Playfair, "The commercial and political atlas and statistical breviary", Cambridge University Press, **(2005)**.
[9] M. Wattenberg and J. Kriss, "Designing for social data analysis", Visualization and Computer Graphics, IEEE Transactions on, vol. 12, no. 4, **(2006)**, pp. 549-557.
[10] J. H. T. Claessen and J. J. Van Wijk, "Flexible linked axes for multivariate data visualization", Visualization and Computer Graphics, IEEE Transactions on, vol. 17, no. 12, **(2011)**, pp. 2310-2316.
[11] C. Shi, W. Cui and S. Liu, "RankExplorer: Visualization of Ranking Changes in Large Time Series Data", Visualization and Computer Graphics, IEEE Transactions on, vol. 18, no. 12, **(2012)**, pp.2669-2678.
[12] J. Kehrer and H. Hauser, "Visualization and visual analysis of multi-faceted scientific data: a survey", IEEE transactions on visualization and computer graphics, IEEE Symposium on. IEEE, vol. 19, no. 3, **(2013)**, pp. 495-513.
[13] T. Saito, H. N. Miyamura and M. Yamamoto, "Two-tone pseudo coloring: Compact visualization for one-dimensional data", Information Visualization, **(2005)**. INFOVIS 2005. IEEE Symposium on. IEEE, **(2005)**, pp. 173-180.
[14] J. Heer, N. Kong and M. Agrawala, "Sizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations", Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, **(2009)**, pp. 1303-1312.
[15] W. S. Cleveland, "Visualizing Data", Hobart Press, **(1993)**.
[16] C. Perin, F. Vernier and J. D. Fekete, "Interactive horizon graphs: improving the compact visualization of multiple time series", Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, **(2013)**, pp. 3217-3226.

Authors

Ruijun Liu, he received his Ph.D. degree in Ecole Centrale de Nantes, France in 2013. He received his M.S. degree in 2009 from Beihang University. He is currently working at Beijing Technology and Business University. Dr. Liu's current research interests include machine learning, virtual reality and 3D reconstruction, etc.

